CORE
Metadata, citation and similar papers at core.ac.uk

感性的情報の推定における単語ユニグラムの有効性の検証

情報科学科メディア情報コース　三浦 弦太　　指導教員：山村 毅

1　はじめに

インターネット上の文章には、SNS での投稿や、通販サイトのレビュー等があり、それぞれの文章には書き手による様々な感性的情報（メッセージ）が込められている。こういった文章から感性的情報を抽出することは極めて有用であり、感性的情報を利用することで、自分が必要としている文章のみを収集することが出来たり、製品の向上に役立つ情報を抽出することができる。ただしこれらのデータは非常に莫大な量であるため、人手で取り扱うのは手間である。

本研究では、感性的情報を自動的に推定する際に、単語ユニグラムがどの程度有効であるかを検証する。実験の対象とするデータには新聞社説を用いる。新聞社説とは、最新のニュースや時事問題を解説すると同時に、評価や主張をする記事のことである。新聞社説から時事問題や国際問題、政策等についての感性的情報を得ることは、メディアの評価や主張を知ることができ、世の中の動きを知る重要な手がかりになると考えられる。

本研究と同じく、新聞社説を題材としてメッセージ情報を抽出する研究には、三浦ら [1] によるもの及び、古田 [2] によるものがある。前者はナイーブベイズ法の多項モデルと多変数ベルヌーイモデルを用いた方法を提案している。一方、後者は k-最近傍決定法とナイーブベイズ法を用いた方法を提案している。

2　分類カテゴリ

本研究で用いる新聞社説では最新のニュースや時事問題に関してのメッセージ性の強い意見文や評価文が多いため、分類カテゴリを以下の 6 つと定めた。

- 肯定 ··· 支持や同意を示しているもの。
  例）人材育成の面からも意味のある活動だ。
- 期待 ··· 良い結果や状態の予期を示しているもの。
  例）互いのトップが指導力を発揮した論戦を期待したい。
- 疑問 ··· 疑念を示しているもの。
  例）国会議員としての誇り、自覚はあるのだろうか。
- 助言 ··· 忠告・アドバイスしているもの。
  例）失敗しても再チャレンジが可能な社会にすべきである。
- 否定 ··· 反対を示しているもの。
  例）事実なら、住民はとても安心して暮らせない。
- 非メッセージ ··· 以上のどれにも当てはまらない。
  例）判決は、同原発の耐震設計に問題があると指摘した。

3　感性的情報の推定手順

3.1　特徴抽出

特徴抽出とは、文章から分類に有効な素性を選択し抽出することである。本研究では、カテゴリの予測に役立つかどうかを表す情報利得値の高い単語 (ユニグラム)n 個 (5~1000) を用いて、各文章から特徴抽出を行なった。

3.2　分類と評価

以下に示す、いずれも確率に基づくパターン分類手法である、ナイーブベイズ法と最大エントロピー法を用いて分類実験を行ない、分類正解率、精度、再現率、F 値、F 値マクロ平均を調べた。評価には 10 分割の交差検定を用いた。

- ナイーブベイズ法
  誤り確率を最小化するという観点 (ベイズ決定則) から、与えられた特徴に対して、条件付き確率を最大にするように、カテゴリを決定する。
- 最大エントロピー法
  与えられた制約を満たすモデルの中で、エントロピーを最大化するように、カテゴリを決定する。

4　実験結果とまとめ

データには、2006 年前半の毎日新聞の社説 344 記事 9492 文を使用する。これは、先行研究 [1] において、既に人手で文単位でカテゴリ分類しているものである。全てのデータを扱うデータセット 1 とデータセット 1 から非メッセージカテゴリを除いたデータセット 2 に対して、ナイーブベイズ法と最大エントロピー法で実験を行なった。表 1 に最大正解率と F 値マクロ平均を示す (カッコ内にはその時の特徴数、データセット 1 は全、データセット 2 はメ、最尤推定による確率 (以下、最尤確率) を用いた時は最、等確率を用いた時は等)。

表 1　全体結果の比較 [%]

	ナイーブベイズ法		最大エントロピー法	
	正解率	F 値マクロ平均	正解率	F 値マクロ平均
全、最	73.9(1000)	57.1(1000)	67.1(5)	19.5(5)
全、等	62.5(1000)	50.5(1000)		
メ、最	74.9(800)	70.9(800)	51.9(5)	27.4(5)
メ、等	73.7(800)	70.3(800)		

最大正解率、F 値マクロ平均共に最も性能が良かったのは、ナイーブベイズ法でデータセット 2(最尤確率) を用いて、特徴数が 800 個の時である。ナイーブベイズ法、最大エントロピー法共に、データセット 1 とデータセット 2 を用いた時の F 値マクロ平均を比較すると、どちらもデータセット 2 を用いた時の方が上回っているため、文章のメッセージを分類するためには、非メッセージカテゴリを除いた上で実験を行なった方が、有効であることがわかる。

また、ナイーブベイズ法と最大エントロピー法の F 値マクロ平均を比較すると、ナイーブベイズ法の方がかなり上回っているため、特徴量に単語ユニグラムを用いた分類実験においては、ナイーブベイズ法の方が有効であると考えられる。

今後は、全てのカテゴリのデータ数を均等にする、単語ユニグラム以外の特徴量を考えるなど、する必要がある。

参考文献

[1] 三浦, 石井, 山村:”ナイーブベイズ分類を用いた新聞社説の感性的情報の分類”, 電気・電子・情報関係学会東海支部連合大会講演論文集,L1-5,2014
[2] 古田麻里絵:”文書からのメッセージ情報の抽出”, 愛知県立大学卒業論文,2009